# **An AUTOMATED SOFTWARE FAILURE PREDICTION technique using hybrid machine learning algorithms**

Roll No. : 21DCS003 Name : Akash Soam

Subject Name : Machine Learning Subject Code : CS-652

Course Component: Class Project

Semester: 7 Year: 2024

Department: Computer Science Institute: NIT Hamirpur

**Datasets :**

1. [**NASA MDP (Metrics Data Program) Datasets**](https://github.com/klainfo/NASADefectDataset) (The datasets include metrics and defect data from NASA software projects.)
2. [**Promise data repository**](http://promise.site.uottawa.ca/SERepository/datasets-page.html) (The PROMISE repository hosts a variety of datasets related to software engineering, including those used for defect prediction.)
3. [**Kaggle Software Defect Prediction dataset**](https://www.kaggle.com/datasets/nazgolnikravesh/software-defect-prediction-dataset)(A dataset available on Kaggle that contains software defect data, which can be used for prediction tasks.)
4. [**SEACRAFT datasets**](https://www.researchgate.net/figure/Dataset-statistics-Data-comes-from-the-SEACRAFT-repository-http-tinycc-seacraft_tbl1_336200370) (This repository includes datasets specifically for software engineering, which include software defect datasets.)
5. [**Eclipse Bug data**](https://download.eclipse.org/scava/aeri_stacktraces/) (Data from the Eclipse project, which includes detailed bug reports and can be used for fault prediction.)

**Project Idea:**

The project aims to develop an advanced software failure prediction technique using hybrid machine learning algorithms. In the proposed approach, the initial phase involves feature selection, which is crucial for improving the prediction accuracy. This phase leverages a genetic algorithm (GA) to optimize the feature set, focusing on identifying the most relevant features from the dataset that significantly impact software failure prediction. The genetic algorithm is enhanced with an improved fitness function to ensure that the selected features contribute effectively to the prediction model.

Once the optimal features are selected, the Decision Tree (DT) algorithm is employed as the classification technique. The Decision Tree is chosen for its interpretability and effectiveness in handling complex datasets. The hybrid GA-DT model is then compared against existing machine learning models, specifically RCSOLDA-RIR and WPA-PSO, which are known for their roles in software failure prediction. The objective is to demonstrate that the proposed hybrid model not only enhances prediction accuracy but also addresses the shortcomings of current models. The successful implementation of this technique could provide a reliable and automated solution for early-stage software failure detection, benefiting industries by reducing the risk of software defects in later stages of development.

**Software/Hardware:**

- Software: Python 3.8+, Scikit-learn, Pandas, Matplotlib, Seaborn, NumPy

- Hardware: Standard desktop/laptop with at least 8GB RAM, Intel i5 Processor/Ryzen 5 processor

**References:**

1. [Predicting software defects using Machine learning techniques](http://wanhussain.com/Pub/Aquil20.pdf)
2. [A comparative study to benchmark cross-project Defect Prediction Approaches](https://ieeexplore.ieee.org/document/7972992)
3. [An ensemble deepboost classifier for software defect prediction](http://refhub.elsevier.com/S2307-1877(23)00002-0/sbref5)
4. [Software defect prediction using stacked denoising autoencoders and two-stage ensemble learning and cuckoo search](http://refhub.elsevier.com/S2307-1877(23)00002-0/sbref11)
5. [High-dimensional hybrid feature selection using interaction information-guided search](http://refhub.elsevier.com/S2307-1877(23)00002-0/sbref8)